



Tools for open data

Filip Hráček
BigClean, Nov 2012

*Big data is **hard**.*

1. Data gathering – hard
2. Data refinement – hard
3. Data analysis – super hard
4. Data sharing – hard

Data gathering

Public Data Explorer

<http://www.google.com/publicdata/directory>





Public Data

Datasets

Metrics

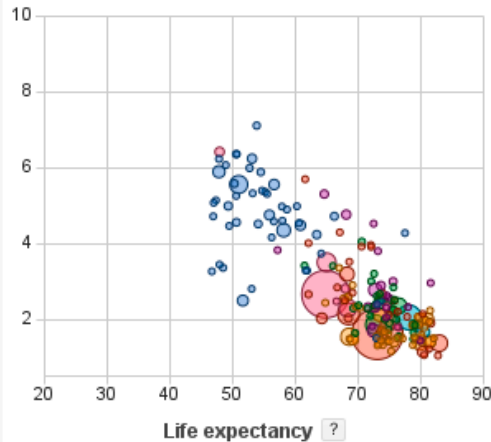
Any data provider (76)

Eurostat (9)
U.S. Census Bureau (5)
U.S. Bureau of Labor
Statistics (3)
World Resources
Institute (2)
Energy Information
Administration (2)

[My Datasets](#)

Fertility rate ?

Countries ?



2009



Living longer with fewer children



This chart correlates life expectancy and number of children per woman for each country in the world. The bubbles are sized by population and colored by region. Over time, most countries have moved towards the bottom right corner of the chart, corresponding to long lives and low fertility. Note the progression of the bubble for China- in the late 60's and 70's life expectancy rose quickly, then the implementation of the one-child policy caused a drop in the number of children per woman.

[Explore the data](#)

Dataset: [World Development Indicators and Global Development Finance](#)
Source: [World Bank](#)

Fusion Tables (Public)

<http://research.google.com/tables>





Tables experimental

Results 1 - 10 of about 209,443 for **czech**. (0.28 seconds)

Web

All Tables

Fusion Tables

Web Tables

Send Feedback

[Born in Europe: Eastern Europe: **Czech Republic**](#)

<https://www.google.com/fusiontables/DataSource?docid=15laXEY8iAD3Y...>

LSOA_CODE	... Czech Republic	geometry	LSOA Boundary
E01000001	3	-0 0952407902109	City of London

[Show more \(569 rows / 4 columns total\)](#) - last modified: Aug 31, 2012

[Czech Fish Subsidy](#)

<https://www.google.com/fusiontables/DataSource?docid...>

Documentation for Czech Fish Subsidy Data This data documentation file records basic information on the ...

Czech Fish Subsidy

[Show more \(512 rows / 16 columns total\)](#) - Fishsubsidy.org - last modified: Aug 1, 2012

[Czech Rep. Vector Outline Copy](#)

<https://www.google.com/fusiontables/DataSource?docid...78>

Czech Rep Vector

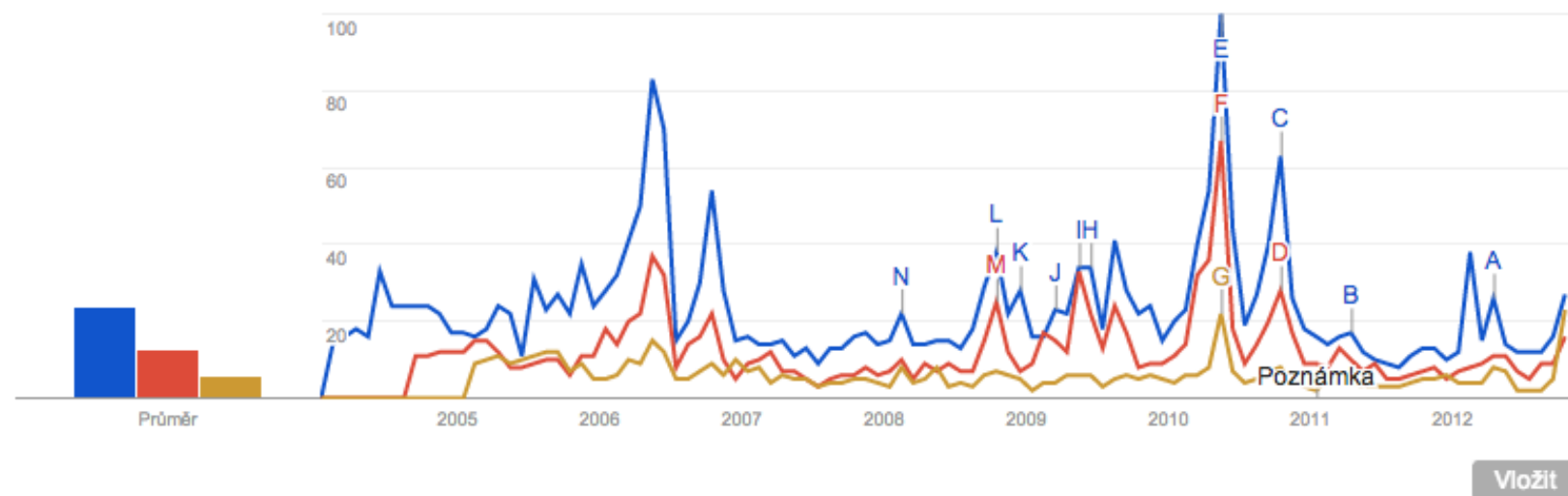
Google Trends

<http://www.google.cz/trends>

Zájem v průběhu času ?

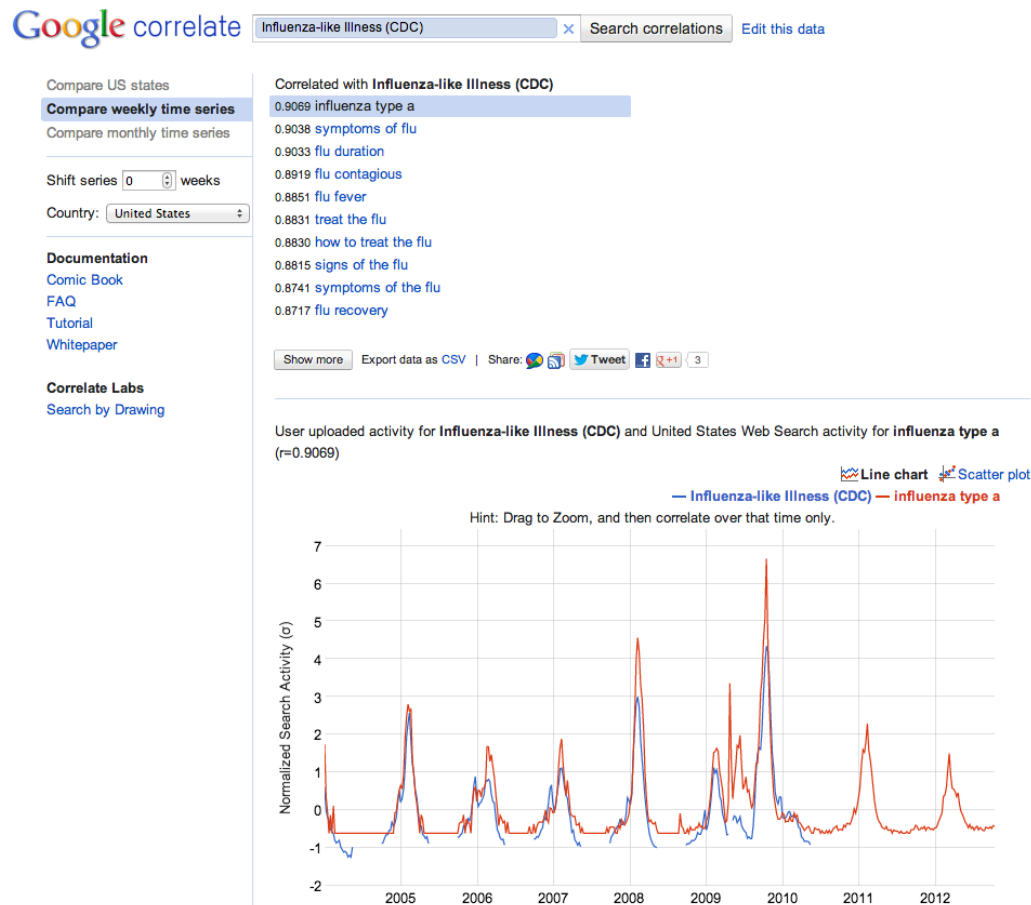
Číslo 100 představuje nejvyšší objem vyhledávání

☒ Titulky zpráv ☐ Předpověď ?



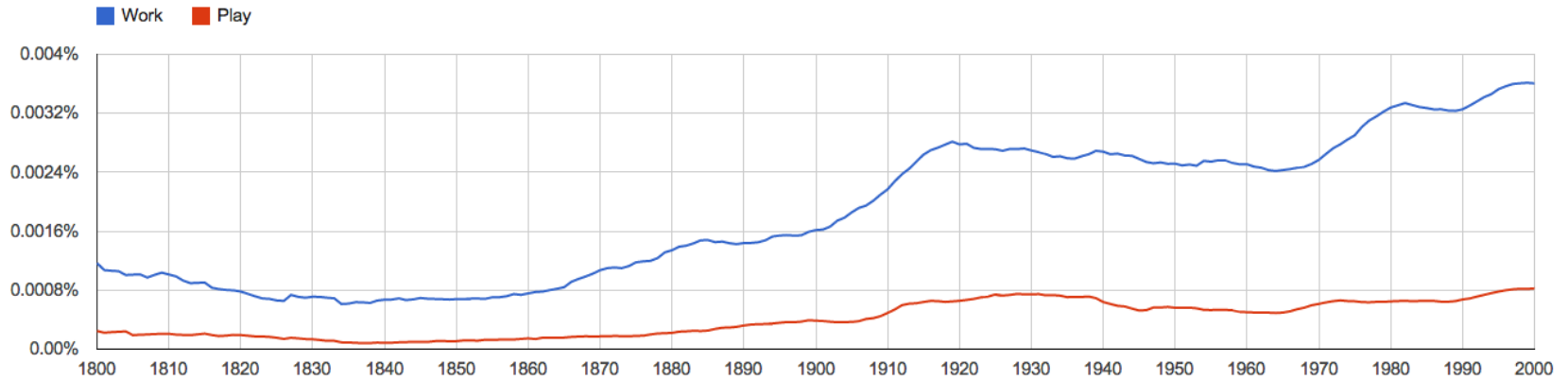
Google Trends Correlate

<http://www.google.com/trends/correlate>



Google Ngram Viewer

<http://books.google.com/ngrams>



Data refinement

Open Refine

<http://code.google.com/p/google-refine/>

Google refine government IT contracts [Permalink](#)

Facet / Filter
Undo / Redo 0

Refresh
Reset All
Remove All

Type of Contract
change invert reset

815 choices Sort by: name count Cluster

FFAA: Fiscal/Financial Agent Agreement 3
FFIP 1
FFP 512
FFP 1
FFP 1
FFP (OPS) 2
FFP (F&E) 1
FFP (Power Supply Retrofit) Old # DTFA01-92-D00004 1
FFP BPA 1
FFP CPAF CPIF 1

512 matching rows (5200 total)

Show as: rows records Show: 5 10 25 50 rows

			Contract ID	Contractor Name	Type of Contract	Date of Award	Start Date	
70.	2038	CGI FEDERAL INCORPORATED	FFP	10/03/2008	10/03/2008	11		
71.	2039	CGI FEDERAL INCORPORATED	FFP	01/09/2009	01/09/2009	09		
72.	2040	CGI FEDERAL	FFP	01/09/2009	01/09/2009	09		
			FFP	03/17/2009	03/23/2009	10		
74.	2042	CGI FEDERAL INCORPORATED	FFP	04/21/2009	04/21/2009	09		
75.	2043	SOLUTIONS ENGINEERING CORP	FFP	11/01/2008	11/01/2008	10		
76.	2044	EVERGREEN INFORMATION TECHNOLO	FFP	11/20/2008	11/20/2008	01		
84.	7946	INTERNATIONAL BUSINESS MACHINES CORPORATION	FFP	10/01/2009	10/01/2009	09		
86.	7947	THE NEWBERRY	FFP	10/01/2009	10/01/2009	09		

Data analysis

Big data.

1. ~~Text editor~~
2. ~~Excel~~
3. ~~Local database~~

Google BigQuery

<https://bigquery.cloud.google.com/>

COMPOSE QUERY

Query History

Job History

API Project

▼ publicdata:samples

github_nested

github_timeline

gsod

nativity

shakespeare

trigrams

wikipedia

New Query

```

1 SELECT
2   year,
3   sum(case when cigarette_use then 1 else 0 end) / count(*) * 1
4 FROM [publicdata:samples.nativity]
5 WHERE year >= 2003
6 GROUP BY year
7 ORDER BY year;

```

RUN QUERY

[Show previous query results](#)

Query Results 3:33pm, 1 Nov 2012

Download as CSV

Save as Table

Chart View

Row	year	percent_dead	
1	2003	8.673828615177587	
2	2004	6.92960535404174	
3	2005	5.942417766803944	
4	2006	5.080331599670038	
5	2007	4.181167102373538	
6	2008	3.331558231942613	

Google Fusion Tables

<http://www.google.com/fusiontables>

<http://research.google.com/tables>

File Edit Tools Help

Rows 1 Cards 1 Map of Geometry

Filter No filters applied

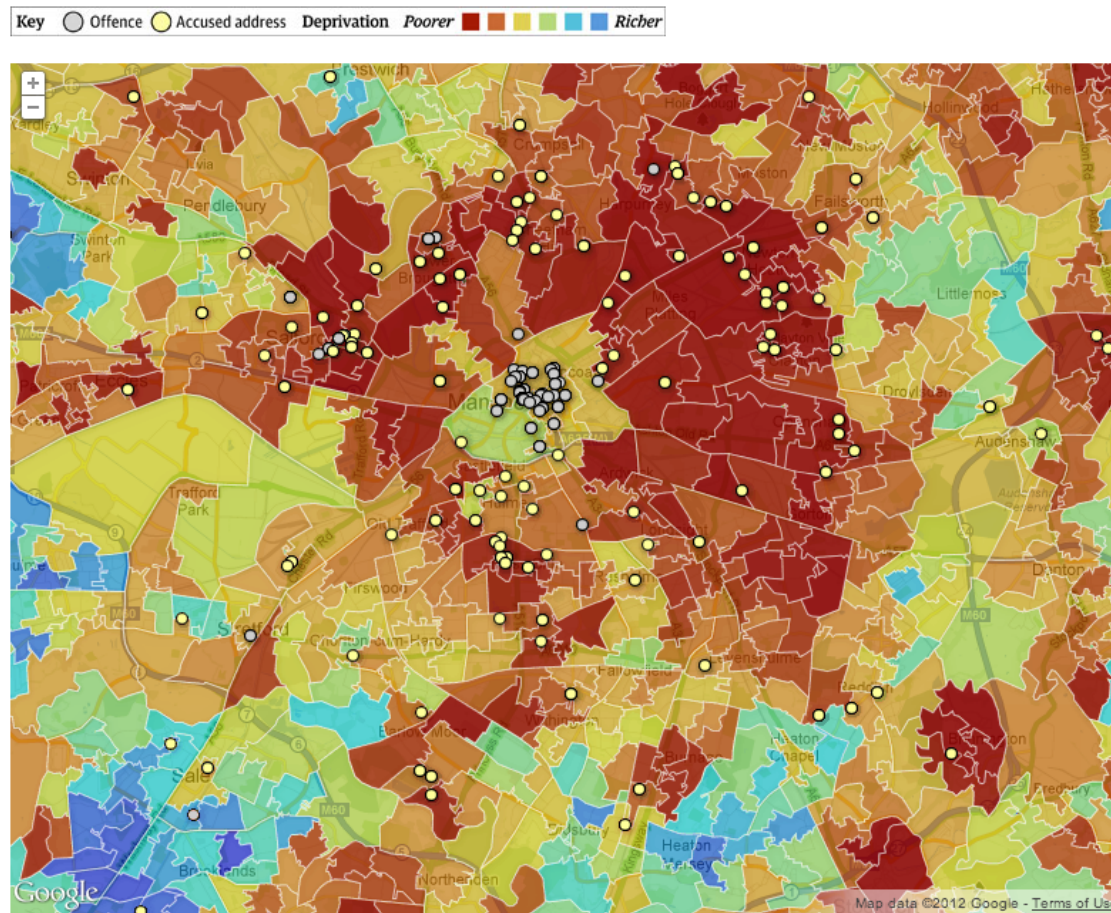
1-85 of 85

Country	rate	homicide	Geometry
Algeria	423.54	1.4	KML...
Argentina	3128.44	5.3	KML...
Armenia	324.2	2.5	KML...
Austria	7078.94	0.7	KML...
Azerbaijan	226.56	2.4	KML...
Bahrain	3785.97	1	KML...
Bangladesh	83.21	2.3	KML...
Belarus	1965.36	8.3	KML...
Belize	3752.61	30.1	KML...
Bermuda	Unknown	1.1	KML...
Bolivia	Unknown	5.3	KML...
Bosnia and Herzegovina	1063.44	1.8	KML...

Data sharing

Google Fusion Tables – in the wild

<http://www.guardian.co.uk/news/datablog/interactive/2011/aug/16/riots-poverty-map>



Google Docs – in the wild

<http://data.blog.ihned.cz/c1-57386250-hledejte-s-nami-fakta-v-projevu-davida-ratha>

Projev Davida Ratha v poslanecké sněmovně 7. září 2012

☆

📁

kdx_2012.42-e ve_12 Debug

filiph@google.com

Comments

Share

File Edit View Insert Help Comment only

🖨

⬆

tak doopravdy, ze to asi bude tak skvar, se kterym se neda nic delat, nepocnopim, ze promotro radši dál už nebudou šetřit, byť jsou tam výpovědi, které říkají, že jistý ministr o tom věděl a dokonce o tom rozhodoval. Byť tam jsou ty výpovědi, tak ony se buď ztratí, nebo se na ně zapomene nebo jsou to nedůvěryhodní svědci. A já myslím, že takhle to půjde v těch dalších nejvyšších kauzách. A pan ministr financí bohužel policii pustí nějakou korunu, aby ten příští rok trochu nějak fungovali. Přece on to dokáže. Konečně někomu tady říkal - copak jsem ti někdy nevyhověl, když jsi chtěl po mně nějaké peníze? - Tuším, že panu poslanci Skokanovi něco takového říkal. - Proč prudiš? - Něco na ten způsob. - Vždyť jsem ti vždycky vyhověl.

Vyhověli panu ministrovi Kubicemu mimochodem. Představte si, že obec Kunice, kde má mimochodem pan ministr Kubice velmi luxusní vilu... On není takový čučkař jako já, který mám řadové domy a ještě je nevlastním sám. On tam má skutečně pěknou vilu za mnoho a mnoho milionů. To víte, plat policejního plukovníka to dovolí si na takovou vilu poctivě vydělat. A jeho syn má vilu hned vedle. Zajímavé je, že to nikoho nezajímá. Mé řadové domy, kde ještě vlastním čtvrtinu, tak ty jsou ve všech médiích a všichni říkají - Rathovy luxusní vily. Nejsou špatné, je to lehce nadprůměrné bydlení, ale žádná... Proti panu Kubicemu, proti panu Kubicemu jsem fakt břídl. Ty policejní platy a platy policejních plukovníků musí být zázračné. Prosím média, jeďte se podívat do Kunice. Podívejte se na vilku pana ministra Kubiceho, na vilku jeho syna a ptejte se, jak si na to jako policejní plukovník vydělal. Poctivec.

10:35 AM Sep 20

Tohle se povedlo, to se ujme.

j.digi.s

2:09 PM Sep 20

Ale ostatní taky kradli!

Patrick Zandl

11:17 AM Sep 25

Tady by to chtělo relevantní informaci z katastru nemovitostí o výměře pozemku a stavby, z toho by se dala udělat představa, nakolik si Kubice žije nad poměry. Mě přišlo ze snímků, že je to pozemek cca 600 metrů plus cca 120 metrů zastavěné plochy vcelku hezkou dřevostavbou, tedy nic, na co by se nedalo našetřit za celoživotní práci. Ovšem na výpisu z katastru by bylo vidět, zda na tom je zástava banky atd. Nevíte někdo přesně to číslo popisné a adresu?

Show less

mjavorek

***Big** thank you!*
(for your attention)

Filip Hráček,
Google Czech Republic

Links

- Public Data Explorer
 - [Import public data form](#)
- Google Correlate
 - [Explanation](#) (Comics)
- Ngram Viewer
 - [Advanced use](#)